Efficiency gains due to using missing data procedures in regression models

Palm, F.C.; Nijman, T.E.

Published in:
Statistical Papers

Publication date:
1988

Statistical Papers 29, 249–256 (1988)

Statistical Papers
Statistische Hefte

Articles

Efficiency gains due to using missing data procedures in regression models

Theo Nijman and Franz Palm

Received: July 20, 1987; revised version: Dec. 21, 1987

The problem of missing observations in regression models is often solved by using imputed values to complete the sample. As an alternative for static models, it has been suggested to limit the analysis to the periods or units for which all relevant variables are observed. The choice of an imputation procedure affects the asymptotic efficiency of the method used to subsequently estimate the parameters of the model. In this note, we show that the relative asymptotic efficiency of three estimators designed to handle incomplete samples depends on parameters that have a straightforward statistical interpretation. In terms of a gain of asymptotic efficiency, the use of these estimators is equivalent to the observation of a percentage of the values which are actually missing. This percentage depends on three R^2-measures only, which can be straightforwardly computed in applied work. Therefore it should be easy in practice to check whether it is worthwhile to use a more elaborate estimator.

We consider the following regression model

$$y_i = \beta x_i + \gamma z_i + e_i , \quad e_i \sim IN(0,\sigma^2), \tag{1}$$

and

$$x_i = \delta z_i + v_i , \quad v_i \sim IN(0,\sigma_v^2), \tag{2}$$

where the regressors x_i and z_i are assumed to be independent of the corresponding disturbances e_i and v_i. The variables y_i and z_i are observed for $i = 1, \ldots, N = N_1 + N_2$, whereas only N_1 values of x_i are observed.

Along with Gouriéroux and Monfort (1981), we assume that the data on x_i are randomly missing; this means that x_i is observed if and only if the random variable ϕ_i takes the value zero. The probability that $\phi_i = 0$ equals $1-\lambda$ and is independent of the parameters in the model. Moreover, the random variables ϕ_i are assumed to be independent of e_j and v_k, for all i, j and k. The probability limit plim $N^{-1}\Sigma z_i^2 = \sigma_z^2$ is assumed to exist and to be finite.

From the strong law of large numbers, N_1N^{-1} converges to $1-\lambda$ with probability one when N goes to infinity. For convenience and without loss of generality, we rearrange the order of the observations such that the subscripts $i = 1,\ldots N_1$ correspond to the observed values of x_i.

Besides the OLS estimator of the regression of y_i on x_i and z_i for $i = 1, \ldots, N_1$ only, denoted by $\hat{\beta}_a$ and $\hat{\gamma}_a$, we consider an estimation procedure in which the missing x_i's are replaced by $\hat{\delta}_a z_i$, where $\hat{\delta}_a$ is the OLS estimate of δ in (2) using the first N_1 observations. This proxy variable for the missing values of x is the optimal prediction of x given the model (2).

The estimate $\hat{\gamma}_{a+b}$ is subsequently computed by OLS on

$$y_i - \hat{\beta}_a\hat{x}_i = \gamma z_i + w_i, \qquad (3)$$

where $w_i = e_i + \phi_i\beta v_i + \phi_i(\beta\delta - \hat{\delta}_a\hat{\beta}_a)z_i$, with $\hat{x}_i = x_i$ and $\phi_i = 0$ if $i \leq N_1$ and $\hat{x}_i = \hat{\delta}_a z_i$ and $\phi_i = 1$ otherwise.

It is straightforward to show that $\hat{\gamma}_{a+b}$ can be alternatively computed by OLS of y_i on $\hat{x}_i$ and z_i

$$y_i = \beta\hat{x}_i + \gamma z_i + \{e_i + \phi_i\beta v_i + \beta\phi_i(\delta - \hat{\delta}_a)z_i\}. \qquad (4)$$

Although the regressors and the disturbance in (4) are not independent, the second moments between $(\hat{x}_i, z_i)$ and the disturbance have zero probability limit as the estimator $\hat{\delta}_a$ converges to δ in probability. Therefore $\hat{\gamma}_{a+b}$

is a consistent estimator of γ. However, the contribution of $\beta\phi_i(\delta - \hat{\delta}_a)z_i$ to the asymptotic variance of $\hat{\gamma}_{a+b}$ is not negligible as plim $N_2N^{-1} = \lambda \neq 0$ for $N \to \infty$.

In the appendix, we show that the large sample distribution of $\hat{\gamma}_{a+b}$ is given by

$$\sqrt{N}(\hat{\gamma}_{a+b} - \gamma) \underset{a}{\sim} N(0,V), \text{ with} \tag{5}$$

$$V = \{(1 - r_{xz}^2)^{-1} + \lambda(\mu^{-1} - 2)\}\ \sigma^2 / (1-\lambda)\sigma_z^2 , \tag{6}$$

where $\mu = \sigma^2(\beta^2\sigma_v^2 + \sigma^2)^{-1}$ and r_{xz}^2 is the theoretical R^2 of the regression (2) of x on $z (r_{xz}^2 < 1)$. The result in (6) has been implicitly obtained by Gouriéroux and Monfort [1981, expression (11) on p. 583].

The relative asymptotic efficiency of $\hat{\gamma}_{a+b}$ with respect to $\hat{\gamma}_a$ is

$$\text{Eff}(\hat{\gamma}_{a+b}) = \text{Avar}(\sqrt{N}\ \hat{\gamma}_{a+b}) / \text{Avar}(\sqrt{N}\ \hat{\gamma}_a) = 1 + \lambda(\mu^{-1} - 2)(1 - r_{xz}^2). \tag{7}$$

According to (7), in large samples using imputed values as in (3) leads to a gain of efficiency compared with using complete observations only if $\mu > \frac{1}{2}$, which is more stringent than the erroneous condition $\mu > (1-\lambda)/(2-\lambda)$ given by Griliches (1986), who also considers the model (1) - (2). Both conditions require that the unpredictable part of x from z is not too important relative to σ^2, the overall noise level of (1).

$$\text{As } \mu = (1 - r_{yxz}^2) / (1 - r_{yz}^2), \tag{8}$$

where r_{yxz}^2 and r_{yz}^2 denote the theoretical R^2's of a regression of y on respectively x and z and on z only $(r_{yz}^2 < 1)$, it is obvious that a sufficient condition for an asymptotic efficiency gain is $r_{yxz}^2 < \frac{1}{2}$, i.e. the

predictible part of y is small.

As noted by Griliches (1986) and others, an asymptotic gain is assured if (4) is estimated by a generalized least squares (GLS) method which takes the correlation structure of the disturbance in (4) into account. Again, the term $\phi_i\beta(\delta - \hat{\delta}_a)z_i$ cannot be neglected (see Palm and Nijman (1982) and Nijman and Palm (1985)). Alternatively, the fully efficient maximum likelihood (ML) estimator can be computed, e.g. using the convenient reparametrisation suggested by Gouriéroux and Monfort (1981). From their results, the relative asymptotic efficiency of the GLS and ML estimators with respect to that of $\hat{\gamma}_a$ can be obtained

$$\text{Eff}(\hat{\gamma}_{GLS}) = 1 - \lambda\mu(1 - r_{xz}^2) \tag{9}$$

and

$$\text{Eff}(\hat{\gamma}_{ML}) = 1 - \lambda\mu(1 - r_{xz}^2) - 2\lambda\mu(1 - \mu)r_{xz}^2. \tag{10}$$

Obviously the GLS and ML estimators are at least as efficient asymptotically as $\hat{\gamma}_a$. A comparison of expression (7) with (9) shows that $\hat{\gamma}_{GLS}$ is at least as efficient as $\hat{\gamma}_{a+b}$ in large samples. Both estimators are equally efficient when $\mu = 1$, that is when conditionally on z, x does not account for the variation of y. From (9) and (10) it follows that the ML estimator is asymptotically more efficient than the GLS estimator. They are equally efficient when $\mu = 1$ or when $r_{xz} = 0$, i.e. when given z, x does not explain y or when z and x are linearly unrelated.

The relative asymptotic efficiency in (7), (9) and (10) only depends on the three magnitudes λ, μ and r_{xz}^2. Equation (9) indicates that in terms of a gain of asymptotic efficiency, the use of GLS is equivalent to the observation of 100 $\mu(1 - r_{xz}^2)$ % of the values of x_i that are actually missing.

Similar expressions can be obtained from (7) and (10) for $\hat{Y}_{a+b}$ and $\hat{Y}_{ML}$ respectively. The values in Table 1 illustrate this result.

Table 1 : Percentage of missing observations that are regained by the use of missing data procedures instead of the complete data only.

$\mu = (1-r^2_{yxz})/(1-r^2_{yz})$	r^2_{xz}	Gain in percentage points for $\hat{Y}_{a+b}$	$\hat{Y}_{GLS}$	$\hat{Y}_{ML}$
.3	.2	-106	24	32
.3	.8	- 27	6	40
.6	.2	27	48	58
.6	.8	7	12	50
.9	.2	71	72	76
.9	.8	18	18	32

Note that a good fit in (2) yielding a "good proxy" for the missing values of x_j does not imply that a large part of the missing information on x_j can be recovered, because of the induced multicollinearity between $\hat{x}_j$ and z_j in (4). Especially, when r^2_{xz} is small, the efficiency gain obtained by using the appropriate estimators can be substantial in large samples. The value of μ is crucial for the asymptotic efficiency of $\hat{Y}_{a+b}$. The loss of efficiency can be important when $\mu < ½$. This loss increases as r^2_{xz} decreases. Finally, if μ is close to one, i.e. x_j is not very important in explaining y in equation (1), all three approaches which take into account the incomplete data, yield about equally efficient estimators in large samples.

To conclude, although it will usually not be possible in more general models to express the asymptotic efficiency in terms of a few magnitudes which can be straightforwardly estimated and used to assess the expected efficiency gain, the ranking of the estimators discussed above holds true

for more general models. For a similar model with aggregate observations on x, we refer to e.g. Palm and Nijman (1982). Results for a dynamic model for y can be found in Nijman and Palm (1985). As OLS based on a proxy variable given an auxiliary equation is not always more efficient than OLS using the complete observations only, we recommend the use of the GLS estimator and whenever possible of the ML method to estimate the parameters of a regression model when some observations on a regressor are missing.

Appendix

In this appendix, we outline the main steps of the proof of the asymptotic properties of the OLS estimator of equation (4), which can be written as

$$N^{\frac{1}{2}} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\gamma}_{a+b} - \gamma \end{pmatrix} = \begin{pmatrix} \frac{X'X}{N} \end{pmatrix}^{-1} \quad \frac{X'u}{N^{\frac{1}{2}}} \quad , \tag{A.1}$$

where X and u are the regressor matrix and the disturbance of (4) respectively.

The inverse of the matrix X'X/N given by

$$\frac{X'X}{N} = \frac{1}{N} \begin{bmatrix} \Sigma x_{1i}^2 + \hat{\delta}_{a_2}^2 \Sigma z_i^2 & \Sigma x_{1i} z_i + \hat{\delta}_a \Sigma z_{2i}^2 \\ \Sigma x_{1i} z_i + \hat{\delta}_a \Sigma z_{2i}^2 & \Sigma z_i^2 \end{bmatrix}, \tag{A.2}$$

where the figures 1 and 2 indicate that we sum over $i = 1,\ldots N_1$ and $i = N_1+1,\ldots N$ respectively (when we sum over all i, no figure is indicated), converges in probability to

$$A^{-1} = (1 - \lambda)^{-1} \sigma_v^{-2} \begin{bmatrix} 1 & -\delta \\ -\delta & \delta^2 + (1-\lambda)\sigma_v^2 \sigma_z^{-2} \end{bmatrix} \tag{A.3}$$

as $\hat{\delta}_a$, $\Sigma x_{1i}^2/N_1$, $\Sigma z_{2i}^2/N_2$, $\Sigma x_{1i}z_{1i}/N_1$, $\Sigma z_{1i}^2/N$, N_1/N converge to δ, $\sigma_x^2 = \delta^2\sigma_z^2 + \sigma_v^2$, σ_z^2, $\delta\sigma_z^2$, σ_z^2 and $1-\lambda$ respectively (these limits are assumed to exist).

The vector $X'u/N^{1/2}$ can be expressed as

$$N^{-1/2}X'u = (N_1/N)^{1/2} \begin{bmatrix} 1 & \delta & \hat{\delta}_a(N_2/N_1)^{1/2} & -\beta\hat{\delta}_a(\Sigma z_{2i}^2)(\Sigma z_{1i}^2)^{-1} & \beta\hat{\delta}_a(N_2/N_1)^{1/2} \\ 0 & 1 & (N_2/N_1)^{1/2} & -\beta(\Sigma z_{2i}^2)(\Sigma z_{1i}^2)^{-1} & \beta(N_2/N_1)^{1/2} \end{bmatrix} \xi, \quad (A.4)$$

where $\xi = [N_1^{-1/2}\Sigma v_{1i}e_{1i}, N_1^{-1/2}\Sigma z_{1i}e_{1i}, N_2^{-1/2}\Sigma z_{2i}e_{2i}, N_1^{-1/2}\Sigma z_{1i}v_{1i}, N_2^{-1/2}\Sigma z_{2i}v_{2i}]'$.

The vector ξ has mean zero and diagonal covariance matrix Δ having $\sigma^2\sigma_v^2$, $\sigma^2\sigma_z^2$, $\sigma^2\sigma_z^2$, $\sigma_v^2\sigma_z^2$ and $\sigma_v^2\sigma_z^2$ on the main diagonal.

From central limit theory for independent random variables, ξ converges in distribution to $N(0,\Delta)$. The matrix premultiplying ξ in (A.4) converges in probability to

$$D = (1-\lambda)^{1/2} \begin{bmatrix} 1 & \delta & \delta[\lambda/(1-\lambda)]^{1/2} & -\beta\delta\lambda/(1-\lambda) & \beta\delta[\lambda/(1-\lambda)]^{1/2} \\ 0 & 1 & [\lambda/(1-\lambda)]^{1/2} & -\beta\lambda/(1-\lambda) & \beta[\lambda/(1-\lambda)]^{1/2} \end{bmatrix}. \quad (A.5)$$

In large samples, the distribution of the OLS estimator is given by

$$N^{1/2}\begin{pmatrix} \hat{\beta} - \beta \\ \hat{Y}_{a+b} - Y \end{pmatrix} \sim N(0,\ A^{-1}D\ \Delta\ D'A^{-1}). \quad (A.6)$$

Notice that $D\Delta D'$ can also be expressed as $D\Delta D' = B + (1-\lambda)\beta^2\ C'\ \Sigma\ C$, where $B = \text{plim}\ N^{-1}\ X'\Omega X$, $C = \text{plim}\ N^{-1}X'W$ and $\Sigma = \sigma_v^2\sigma_z^{-2}$ is the asymptotic variance of $N_1^{1/2}\hat{\delta}_a$, with Ω being a diagonal matrix with typical element $\sigma^2 + \phi_i\beta^2\sigma_v^2$, W being a vector with typical element $\phi_i z_i$. This finding shows that the

term $\beta\phi_i(\delta-\hat{\delta}_a)z_i$ in the disturbance of (4) contributes to the variance of the regression coefficient estimates and therefore cannot be ignored.

The asymptotic variance of $N^{\frac{1}{2}}\hat{\gamma}_{a+b}$ is the second element of the main diagonal of $A^{-1}D\Delta D'A^{-1}$

$$\text{Avar}(N^{\frac{1}{2}}\hat{\gamma}_{a+b}) = \delta^2\sigma^2\sigma_v^{-2}(1-\lambda)^{-1} + \sigma^2\sigma_z^{-2} + \lambda(1-\lambda)^{-1}\beta^2\sigma_v^2\sigma_z^{-2}. \qquad (A.7)$$

Some straightforward algebra yields the result in equation (6).

Acknowledgement

The authors thank Professor Z. Griliches and an unknown referee for their comments on an earlier version of this note.

References

Gouriéroux, C., and A. Monfort (1981), "On the problem of missing data in linear models", Review of Economic Studies, 48, 579-586.

Griliches, Z. (1986), "Economic data issues", in Z. Griliches and M.D. Intriligator, eds, Handbook of Econometrics, North-Holland, Amsterdam, 1466-1514.

Nijman, Th.E., and F.C. Palm (1985), "Consistent estimation of a regression model with incompletely observed exogenous variable", Netherlands Central Bureau of Statistics, unpublished paper.

Palm, F.C., and Th.E. Nijman (1982), "Linear regression using both temporally aggregated and temporally disaggregated data", Journal of Econometrics, 19, 333-343.

Nijman Th.E.
Department of Econometrics
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

Palm F.C.
Department of Economics
University of Limburg
P.O. Box 616
6200 MD Maastricht
The Netherlands